

# Distribution of Sample Proportions and Confidence Interval for a Population Proportion

Shannon Pileggi

STAT 217

# OUTLINE

Overview

Simulation

Distribution of a Sample Proportions

Confidence Interval in General

Confidence Interval for a Population Proportion

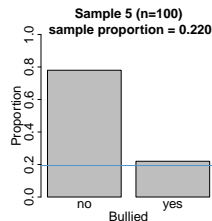
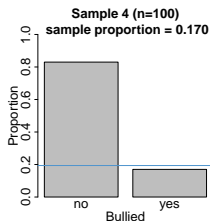
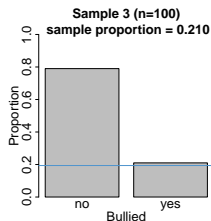
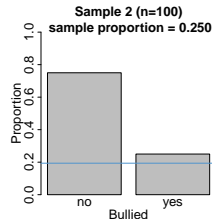
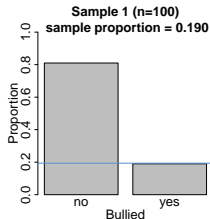
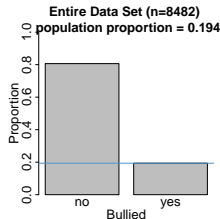
Understanding the CI

# The Data

From the CDC's 2013 Youth Risk Behavior Surveillance System

|      | gender | height_m | weight_kg | bmi  | carried_weapon | bullied | days_drink |
|------|--------|----------|-----------|------|----------------|---------|------------|
| 1    | female | 1.73     | 84.37     | 28.2 | yes            | no      | 30         |
| 2    | female | 1.6      | 55.79     | 21.8 | no             | yes     | 1          |
| 3    | female | 1.5      | 46.72     | 20.8 | no             | yes     | 0          |
| 4    | female | 1.57     | 67.13     | 27.2 | no             | yes     | 0          |
| 5    | female | 1.68     | 69.85     | 24.7 | no             | no      | 0          |
| 6    | female | 1.65     | 66.68     | 24.5 | no             | no      | 1          |
| 7    | male   | 1.85     | 74.39     | 21.7 | no             | no      | 0          |
| 8    | male   | 1.78     | 70.31     | 22.2 | yes            | no      | 0          |
| 9    | male   | 1.73     | 73.48     | 24.6 | no             | yes     | 0          |
| 10   | male   | 1.83     | 67.59     | 20.2 | no             | no      | 0          |
| ⋮    | ⋮      | ⋮        | ⋮         | ⋮    | ⋮              | ⋮       | ⋮          |
| 8482 | male   | 1.73     | 68.95     | 23   | no             | no      | 0          |

# The idea



## Parameters vs statistics

A **parameter** is a numerical summary of a population. It is usually *unknown*, although we can make *assumptions* about parameter values for population distributions. We generally use Greek letters (without bars or hats) to denote population parameters:

|                               |          |
|-------------------------------|----------|
| population mean               | $\mu$    |
| population standard deviation | $\sigma$ |
| population proportion         | $p$      |

A **statistic** is a numerical summary of the sample. It is estimated from observed data. We generally use lower case letters (with bars or hats) to denote sample statistics:

|                           |           |
|---------------------------|-----------|
| sample mean               | $\bar{x}$ |
| sample standard deviation | $s$       |
| sample proportion         | $\hat{p}$ |

# Describing distributions

## Group Exercise

When describing distributions, what are the three features you should address?

1.

2.

3.

## Three distributions to keep in mind

1. The **population distribution** refers to the actual distribution of a variable in a population.
2. The **data distribution** refers to the distribution of observed values from a *single* sample.
3. The **sampling distribution** refers to the distribution of a statistic from *many* samples.

sampling distribution of the sample proportion = the distribution of sample proportions ( $\hat{p}$ ) from many samples

sampling distribution of the sample means = the distribution of sample means ( $\bar{x}$ ) from many samples

## Overview

## Simulation

## Distribution of a Sample Proportions

## Confidence Interval in General

## Confidence Interval for a Population Proportion

## Understanding the CI



## Population and data distribution of bullied

Consider the 8,482 observations from the YRBSS data set to be the *entire* population of interest. Now let's describe the **population distribution** of bullied.

- ▶ True population proportion:

## Example data distributions from bullied

Now let's take three random samples of size  $n = 10$  from the population distribution of bullied. Each random sample represents a **data distribution**.

|          | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $\hat{p}$ |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|-----------|
| Sample 1 |       |       |       |       |       |       |       |       |       |          |           |
| Sample 2 |       |       |       |       |       |       |       |       |       |          |           |
| Sample 3 |       |       |       |       |       |       |       |       |       |          |           |

## Many samples from bullied

Let's repeat the process and take 1000 random samples of size  $n = 10$  from the population distribution of bullied.

|            | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $\hat{p}$ |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|-----------|
| Sample1    | no    | yes   | no    | no    | no    | no    | no    | no    | no    | yes      | 0.2       |
| Sample2    | no    | no    | no    | no    | no    | no    | no    | no    | no    | yes      | 0.1       |
| Sample3    | no    | no    | yes   | no    | no    | yes   | yes   | no    | no    | no       | 0.3       |
| Sample4    | no    | no    | no    | yes   | yes   | no    | no    | no    | yes   | no       | 0.3       |
| Sample5    | no    | yes   | no    | no    | no    | no    | no    | yes   | no    | no       | 0.2       |
| Sample6    | no    | n     | no    | yes   | yes   | yes   | yes   | no    | no    | no       | 0.4       |
| Sample7    | yes   | no    | no    | no    | no    | no    | no    | no    | no    | yes      | 0.2       |
| Sample8    | no    | yes   | yes   | yes   | no    | no    | no    | no    | no    | yes      | 0.4       |
| Sample9    | no    | yes   | no    | no    | no    | no    | no    | yes   | no    | no       | 0.2       |
| Sample10   | no    | yes   | no    | no    | no    | no    | no    | no    | no    | no       | 0.1       |
| Sample11   | yes   | no    | no    | no    | no    | no    | no    | no    | no    | no       | 0.1       |
| ⋮          | ⋮     | ⋮     | ⋮     | ⋮     | ⋮     | ⋮     | ⋮     | ⋮     | ⋮     | ⋮        | ⋮         |
| Sample1000 | no    | no    | no    | no    | no    | no    | no    | no    | no    | no       | 0         |

## Group Exercise

What do you think will be the shape of the distribution of the 1000 sample proportions?

1. bell-shaped
2. left-skewed
3. right-skewed
4. uniform

## Simulated sampling distribution, example 1

The collection of the sample proportions from the 1000 samples of size  $n = 10$  represents a simulated **sampling distribution** of the sample proportion.

- ▶ Shape of the sampling distribution:
- ▶ Mean of the sampling distribution:
- ▶ Standard deviation of the sampling distribution:

# Re-cap, example 1

## Population distribution

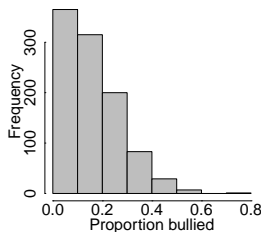
$$p = 0.194$$

## Single data distribution ( $n = 10$ )

no yes no no no  
no no no no yes

$$\hat{p} = 0.2$$

## Distribution of 1000 sample proportions from samples of size $n = 10$



$$\text{mean} = 0.202$$

$$\text{sd} = 0.127$$

## Group Exercise

What do you think will happen to the distribution of sample proportions if we increase the sample size for each individual sample from  $n = 10$  to  $n = 100$ ? (The number of samples will stay the same at 1000.)

The shape will be \_\_\_\_\_, the mean will \_\_\_\_\_, the standard deviation will \_\_\_\_\_.

1. shape: right-skewed, left-skewed, approximately normal
2. mean: increase, decrease, remain the same
3. standard deviation: increase, decrease, remain the same

## Simulated sampling distribution, example 2

The collection of the sample proportions from the 1000 samples of size  $n = 100$  represents a simulated **sampling distribution** of the sample proportion.

- ▶ Shape of the sampling distribution:
- ▶ Mean of the sampling distribution:
- ▶ Standard deviation of the sampling distribution:



## Re-cap, example 2

### Population distribution

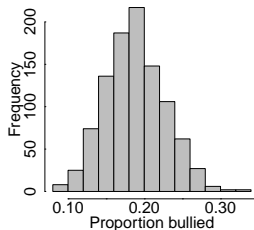
$$p = 0.194$$

### Single data distribution ( $n = 100$ )

no no no no no no  
no yes no no yes  
no no no no no no  
no no no no no no  
no no no yes no  
no no no no no no  
no no no no no no  
yes ...

$$\hat{p} = 0.173$$

### Distribution of 1000 sample proportions from samples of size $n = 100$



$$\begin{aligned}\text{mean} &= 0.193 \\ \text{sd} &= 0.04\end{aligned}$$

# Summary

---

| Feature | Example 1 ( $n = 10$ ) | Example 2 ( $n = 100$ ) |
|---------|------------------------|-------------------------|
|---------|------------------------|-------------------------|

---

*Observed in simulation*

Shape

Mean

Std Dev

---

*According to theory*

Shape

Mean

Std Dev

---

Overview

Simulation

Distribution of a Sample Proportions

Confidence Interval in General

Confidence Interval for a Population Proportion

Understanding the CI

# Distribution of Sample Proportions

OR: the sampling distribution of the sample proportion

For a random sample of size  $n$  from a population with population proportion  $p$ , the distribution of sample proportions has

$$\text{mean} = p, \text{ standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

Saying the same thing, but with more notation:

$$\text{mean}(\hat{p}) = p, \text{ sd}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

# Distribution of Sample Proportions

OR: the sampling distribution of the sample proportion

When  $np \geq 10$  and  $n(1 - p) \geq 10$ , then the the distribution of sample proportions has an **approximately normal** shape. That is, when this condition is satisfied, the distribution of sample proportions has:

- ▶ shape = normal
- ▶ mean =  $p$
- ▶ standard deviation =  $\sqrt{\frac{p(1 - p)}{n}}$

The condition  $np \geq 10$  and  $n(1 - p) \geq 10$  means that you have at least 10 expected 'successes' and 10 expected 'failures.'

## Group Exercise

Suppose that 80% of Americans prefer milk chocolate to dark chocolate. For which of the following sample sizes would the distribution of the sample proportions of Americans that prefers milk chocolate be approximately normally distributed? Mark all that apply.

1.  $n = 20$
2.  $n = 40$
3.  $n = 60$
4.  $n = 80$

## Group Exercise

Which of the following affects the variability (or spread) in the distribution of the sample proportions? Select all that apply

1. the population mean
2. the population standard deviation
3. the sample size
4. the number of samples collected

## Group Exercise

When discussing sampling distributions, what started off as KNOWN and UNKNOWN?

- A. population proportion: (1) known OR (2) unknown
- B. sample proportion: (1) known OR (2) unknown



## Example

Suppose that 80% of Cal Poly students own a Mac laptop, and that we take samples of size  $n = 100$  from the population of all Cal Poly students. Specify (with justification) the following features of the distribution of sample proportions and sketch the distribution of sample proportions.

1. shape
2. mean
3. std dev

Overview

Simulation

Distribution of a Sample Proportions

Confidence Interval in General

Confidence Interval for a Population Proportion

Understanding the CI

# Estimation

Given that we generally collect data from a sample (and not a population), how do we *estimate* population parameter values reliably? Examples of estimation:

- ▶ A **study** found that the average breath alcohol concentration (BrAC) was .091 when subjects drank alcohol mixed with a diet drink. By comparison, BrAC was .077 when the same subjects consumed the same amount of alcohol but with a sugary soda.
- ▶ **Researchers** estimate that domestic cats are responsible for the deaths of between 1.4 and 3.7 billion birds and 6.9-20.7 billion mammals annually.

## Point estimate vs interval estimate

A **point estimate** is a **single number** that is our 'best guess' for the population parameter. Point estimates are given by sample statistics.

- ▶ the average number of hours of sleep on a typical night is 7.024 ( $\bar{x} = 7.024$ )
- ▶ 100/208 students indicated the Emory was their 1<sup>st</sup> choice ( $\hat{p} = 0.48$ )

An **interval estimate** is an **interval of numbers** within which the population parameter value is believed to fall.

- ▶ What is a plausible range for the **population mean** ( $\mu$ ) number of hours of sleep of Emory students in a typical night?
- ▶ What is a plausible range for the **population proportion** ( $p$ ) of students for whom Emory was their 1<sup>st</sup> choice?

# Confidence Interval

A **confidence interval** is an interval containing the most believable values for a population parameter.

- ▶ The probability that this method produces an interval that captures the true parameter value is called the **confidence level**.
- ▶ The confidence level is a number close to 1, and is most commonly 0.95.
- ▶ A confidence interval with a confidence level of 0.95 is called a 95% confidence interval.

## General Form of Confidence Interval:

point estimate  $\pm$  margin of error

- ▶ the **point estimate** is your best guess of a population parameter, like  $\bar{x}$  or  $\hat{p}$
- ▶ the **margin of error** measures how accurate the point estimate is likely to be in estimated a parameter

# Relationship between sampling distribution and confidence intervals

|                       | <b>Sampling<br/>Distribution</b> | <b>Confidence<br/>Interval</b> |
|-----------------------|----------------------------------|--------------------------------|
| Population proportion | known                            | unknown                        |
| Sample proportion     | unknown                          | known                          |

Overview

Simulation

Distribution of a Sample Proportions

Confidence Interval in General

Confidence Interval for a Population Proportion

Understanding the CI



## The research question

How can we estimate the population proportion of Cal Poly students that own a Mac laptop?

In random sample of Cal Poly students, 50 out of 67 students reported that they own a Mac laptop.

## CI for a population proportion

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

*point estimate  $\pm$  critical value  $\times$  standard error*

*point estimate  $\pm$  margin of error*

- ▶ the **point estimate** is your best guess of a population parameter  $\rightarrow \hat{p}$
- ▶ the **critical value** establishes your degree of confidence for that interval  $\rightarrow$  use  $z$
- ▶ the **standard error** allows for uncertainty in that point estimate  $\rightarrow \sqrt{\hat{p}(1 - \hat{p})/n}$
- ▶ the **margin of error** is the (critical value  $\times$  standard error), and is everything after the  $\pm \rightarrow z \times \sqrt{\hat{p}(1 - \hat{p})/n}$

## Choice of *approximate* $z$

point estimate  $\pm z^* \times se \rightarrow$  general form of CI

point estimate  $\pm 1 \times se \rightarrow$  approximate \_\_\_\_ CI

point estimate  $\pm 2 \times se \rightarrow$  approximate 95% CI

point estimate  $\pm 3 \times se \rightarrow$  approximate \_\_\_\_ CI

# Elements of an interpretation of a confidence interval

1. State the confidence level
2. Refer to the population
3. State the parameter being estimated
4. Utilize context
5. Include a range of values

At the 1 % confidence level, we estimate that the 2  
3 of 4 is in the interval 5.

## Evaluating claims

Different faculty members have different guesses on the percent of all Cal Poly students that own laptop. Based on the interval calculated, which of these claims are plausible? Mark all that apply.

1. Professor A claims that 60% of students own a Mac laptop.
2. Professor B claims that 70% of students own a Mac laptop.
3. Professor C claims that 80% of students own a Mac laptop.
4. Professor D claims that 90% of students own a Mac laptop.

## Conditions required for a CI for $p$

1. The observations are independent. [◀ Discussion](#)
2.  $n\hat{p} \geq 10$  and  $n(1 - \hat{p}) \geq 10$   
(at least 10 observed “successes” and 10 observed “failures”)

# Steps to constructing a confidence interval for a population proportion

1. Check your conditions.
2. Identify  $z^*$  for your specified level of confidence.

|                  |      |      |      |      |
|------------------|------|------|------|------|
| Confidence level | 80%  | 90%  | 95%  | 99%  |
| $z^*$            | 1.28 | 1.65 | 1.96 | 2.58 |

3. Calculate the interval:  $\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

## Example

In random sample of Cal Poly students, 50 out of 67 students reported that they own a Mac laptop. Compute and interpret a 95% CI for the population proportion of Cal Poly students that own a Mac laptop.



## Overview

## Simulation

## Distribution of a Sample Proportions

## Confidence Interval in General

## Confidence Interval for a Population Proportion

## Understanding the CI

## Group Exercise

A 95% confidence interval for the true proportion of US citizens who are opposed to issuing traffic tickets from traffic cameras is (0.57, 0.63) based on a sample of 1000 individuals.

What is the point estimate for the proportion of sampled individuals who are opposed to issuing traffic tickets from traffic cameras?

1. 0.57
2. 0.63
3. 0.60
4. 0.95
5. not enough information to determine

## Group Exercise

$$\hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \text{ 90\% } z^* = \underline{\hspace{2cm}}, \text{ 95\% } z^* = \underline{\hspace{2cm}}$$

The value of  $z$  for a 95% CI is \_\_\_\_\_ than the value of  $z^*$  for a 90% CI. This means that higher confidence levels correspond to \_\_\_\_\_ confidence intervals.

1. greater; wider
2. greater; narrower
3. less; narrower
4. less; wider

What factors affect the width of the CI? Constructing a confidence interval is a *compromise* between an acceptable width of your confidence interval and the desired level of confidence in correct inference.

## Group Exercise

Will a 95% confidence interval always contain the **estimate** of the population proportion ( $\hat{p}$ )?

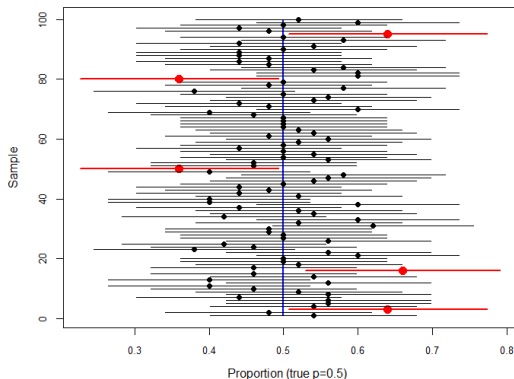
1. Yes
2. No

Will a 95% confidence interval always contain the **population** proportion ( $p$ )?

1. Yes
2. No

## The meaning of a 95% CI

If we were to repeatedly sample over and over again, *in the long run* 95% of our confidence intervals would make correct inference. That is, if we took 100 random samples and calculated 100 confidence intervals, we would expect 95 of those confidence intervals to capture the true parameter value.



This figure shows 100 samples of size  $n = 50$  where the true  $p = 0.50$ . Note that 5 of the 100 CI's don't actually capture the true  $p$ .

## The meaning of a 95% CI, continued

- ▶ 95% of samples of this size will produce confidence intervals that capture the true proportion.
- ▶ This can be longwinded, so we say “We are 95% confident that the true proportion lies in this interval.”
- ▶ In reality, we cannot know whether or our sample is one of the 95% that captured  $p$ , or one of the unlucky 5% that did not catch  $p$ .
- ▶ 95% confident means that we arrived at this interval by a method that gives us correct results 95% of the time.

## Group Exercise

Based on data from the Winter 2016 STAT 217 class, a 95% confidence interval for the proportion of Cal Poly students who own a Mac laptop is 0.64 to 0.85.

Which of the following is a correct interpretation?

1. We are 95% confident that the proportion of Cal Poly students from this sample who own a Mac laptop is between 0.64 and 0.85.
2. We are 95% confident that the population proportion of Cal Poly students who own a Mac laptop is between 0.64 and 0.85.
3. 95% of the time the proportion of Cal Poly students who own a Mac laptop is between 0.64 and 0.85.
4. More than one statement is correct.

## Interpreting the CI

- ▶ Interpreting a CI can be challenging
- ▶ Students often try to use their own words, and get the interpretation incorrect (just use my words).
- ▶ Confidence intervals are about values of population parameters, so *both* pieces of this information **must** be included in the interpretation.



## Group Exercise

| Sample 1         | Sample 2         |
|------------------|------------------|
| $\hat{p} = 0.35$ | $\hat{p} = 0.35$ |
| $n = 50$         | $n = 100$        |

Suppose we construct a 95% confidence interval for  $p$  for both samples. How will the confidence intervals compare?

1. The width of the intervals from the two samples will be the same.
2. The CI for sample 1 will be wider than the CI for sample 2.
3. The CI for sample 2 will be wider than the CI for sample 1.
4. There is not enough information to determine.

## Group Exercise

For each of the following scenarios, indicate if the research question should be answered with a confidence interval for a **mean** or **proportion**. We want to know something about...

1. how many times a day a baby laughs
2. whether or not babies are born with blue eyes
3. the number of months until the baby first sleeps through the night
4. if babies can sit up independently by age 6 months
5. the heart rate of baby immediately after birth

## When are observations not independent?

Observations are not independent when they are correlated with each other. This occurs when observations are related and are more similar to each other than other observations in the data set.

- ▶ Measurements made on the same subject are typically not independent.
  - ▶ circumference of right and left thigh
  - ▶ blood pressure before and after a treatment
- ▶ Observations from different subjects may not always be independent.
  - ▶ siblings, twin pairs, husband-wife
- ▶ Observations from groups of subjects may not always be independent.
  - ▶ children that attend the same elementary school, patients that attend the same in-town clinic